CGEMS Prostate Cancer Phase I Build 2.0

11/29/2009

1 Introduction

CGEMS Prostate Cancer Phase I build 1.0 was released on the caBIG web portal (https://caintegrator.nci.nih.gov/cgems) in early 2007. The released data set was comprised of more than 1,100 prostate cancer cases and an equivalent number of controls from the Prostate, Lung, Colon and Ovarian Prevention study (PLCO - http://prevention.cancer.gov/programs-resources/groups/ed/programs/plco). The cases and controls were genotyped on both HumanHap300 and HumanHap240 platforms; together yielding data on nearly 550,000 SNPs. Build 1.0 of the data included genotypes, phenotypes, frequencies and genotype-phenotype association results, accessible by application and approval by the CGEMS Data Access Committee at the National Cancer Institute (NCI). We invite readers to refer to Yeager *et al.* (2007)[1], especially the Supplementary Methods (http://www.nature.com/ng/journal/v39/n5/extref/ng2022-S5.pdf), for a thorough description of Build 1.0.

Build 1.0 was the first GWAS conducted at CGF, and has resulted in numerous scientific publications with major impact on our understanding of the genetic determinants of prostate cancer risk and validating the GWAS design and analytic methodologies. Over the past few years, as more genome wide scans have been analyzed at the Core Genotyping Facility of the NCI, the tools and methodologies have significantly improved. To meet the demand for high quality data sharing within the scientific community, we release CGEMS Prostate Cancer Phase I build 2.0, which incorporates many improvements, including a refined quality control process, minor updates to subject phenotype, and fitting of a wider selection of association models to enable better comparisons with others' results.

2 Genotype Quality Control

Genotypes for build 1.0 were extracted from the Genotype Final Reports with genotypes called by Illumina Inc. for the original CGEMS analysis. In contrast, build 2.0 started with the raw array intensity files (IDAT files), which were clustered and genotypes called with the most recent version of Illumina's Genome Studio software V2009.1. Prior to clustering, the project sample sheets that define the mapping between samples and assay barcodes were updated to correct a previously identified plate transposition (this was also corrected in Build 1.0), data from ineligible samples were removed, and low performing samples were temporarily excluded (<98% competition based on a previous clustering). Genotype calling was performed on all samples including those excluded for low performance. The resulting genotypes between build 1.0 and build 2.0 are highly concordant, approaching 100%, based on the informative comparison (detailed in Section 3).

A total of 2,544 samples were genotyped on the 317K platform, and a total of 2,426 samples on the 240K platform. After excluding ineligible samples, CEPH QC samples, and samples with all phenotype data missing, 2,321 samples on the 317K platform and 2,301 samples on the 240K platform were analyzed. The genotyped samples are summarized in Table 1a and a detailed breakdown of individuals by case/control status is summarized in Table 1b.

Table 1a. Summary of genotyped samples

Number of samples	**Illumina 317K**	**Illumina 240K**
Genotyped samples	2,544	2,426
Ineligible samples	23	18
CEPH QC samples	176	96
Missing phenotype	24	11
Remaining samples	2,321	2,301

Table 1b. Summary of genotyped subjects

Number of subjects	**Phenotype**		
	Cases	**Controls**	**Total**
Illumina 317K	1,164	1,109	2,273
Illumina 240K	1,151	1,102	2,253

2.1. Completion Rates

Genotype completion rates, the proportion of non-missing to total possible genotypes, are summarized in Table 2. The distributions of completion rate by sample and by locus are shown in Figures 1a and1b. A brief summary of the sample/locus counts at 100th, 99th, 95th, 90th and 50th quantiles are provided as insert in each figure. Samples with the completion rates lower than 94.5% were excluded from the final analytic data set.

Table 2. Completion rates by QC group

	Completion rate (%)		
	All	**Informative***	**Uncalled loci**
Illumina 317K	98.91	99.48	1,826
Illumina 240K	98.84	99.72	2,145

* Informative indicates that all uncalled loci – those loci with no genotypes called – are excluded from the determination of the expected number of non-missing genotypes.

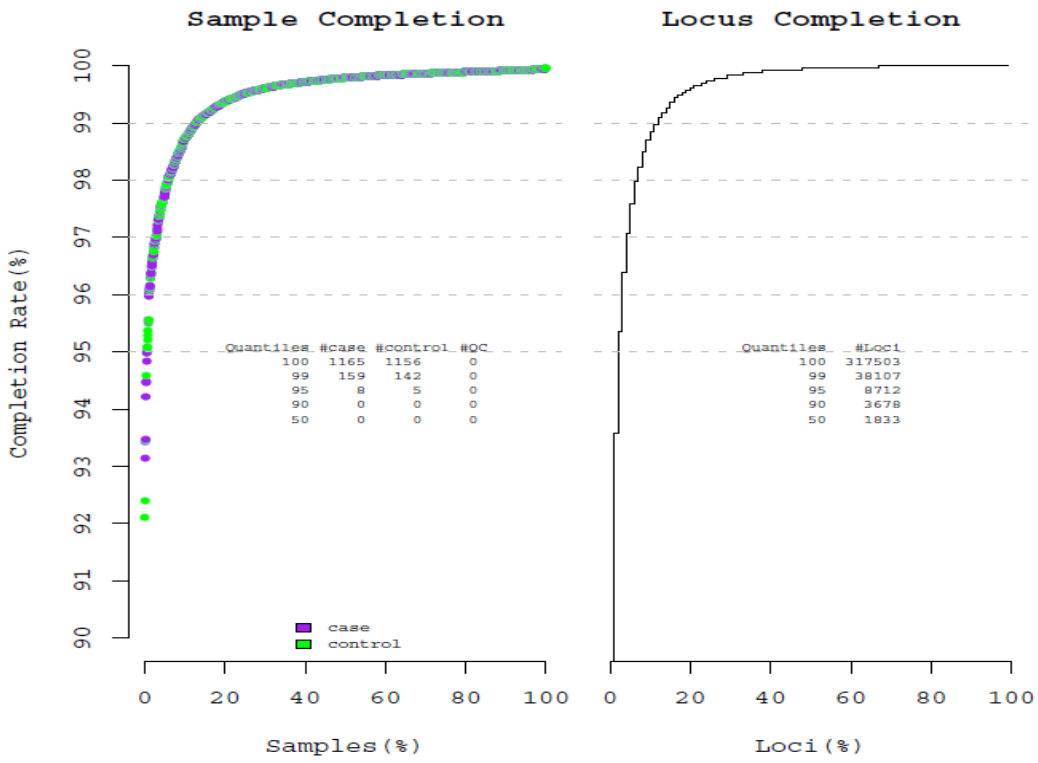

Figure 1a. Illumina 317K completion by sample (left) and by locus (right).

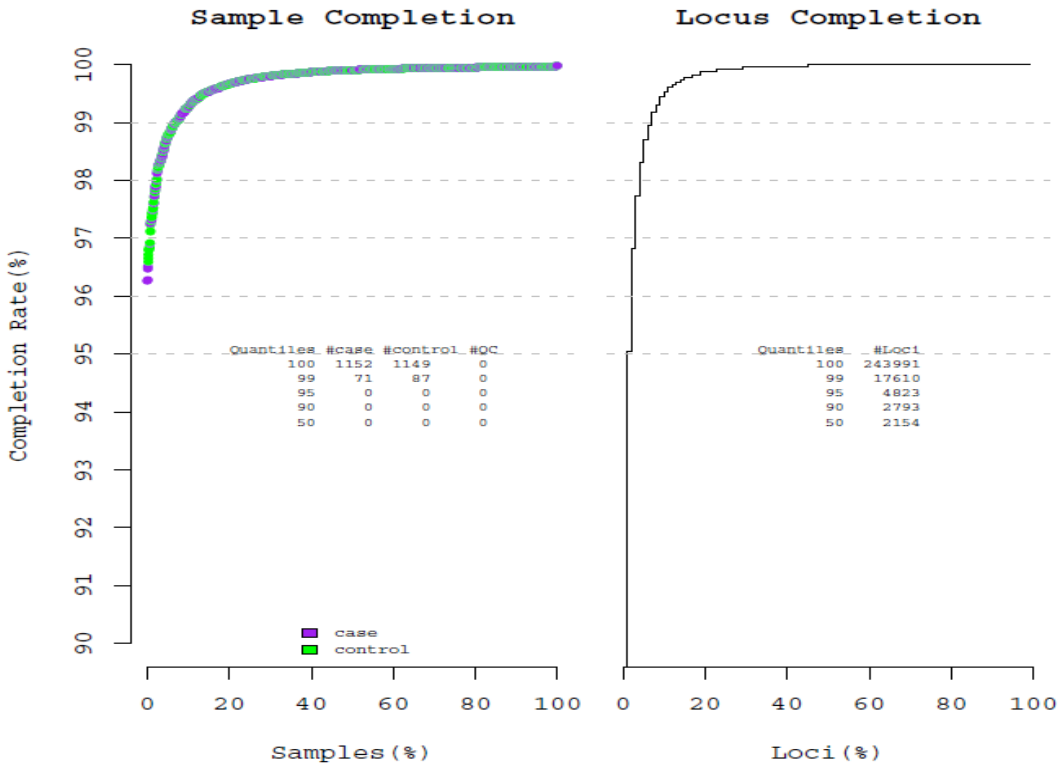

Figure 1b. Illumina 240K completion by sample (left) and by locus (right).

2.2 Sample Heterozygosity

The presence of too few or too many heterozygote genotypes is often indicative of large chromosomal abnormalities and is a good indicator of problematic samples and assays. Thus the mean heterozygosity for each sample for each assay was computed based on only autosomal SNPs in order to detect outliers. The distribution of the sample heterozygosity and its relationship to sample completion rate are shown in Figure 2a and 2b. The mean sample heterozygosity varies by population and assay content; for the PLCO study on Illumina 317K it is around 34%, and that on the 240K it is 29%. Samples with heterozygosity higher than 36% for 317K data or higher than 35% for 240K data were removed from the final analytic data set.

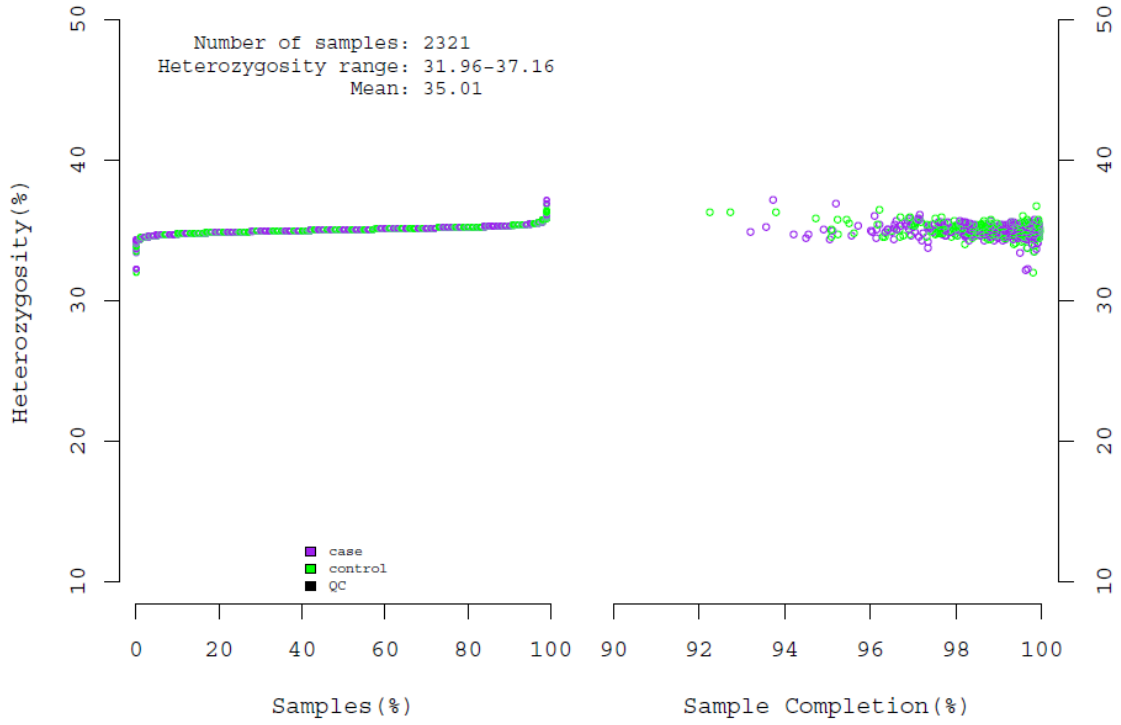

Figure 2a. Illumina 317K heterozygosity distribution (left) and its relation to completion (right).

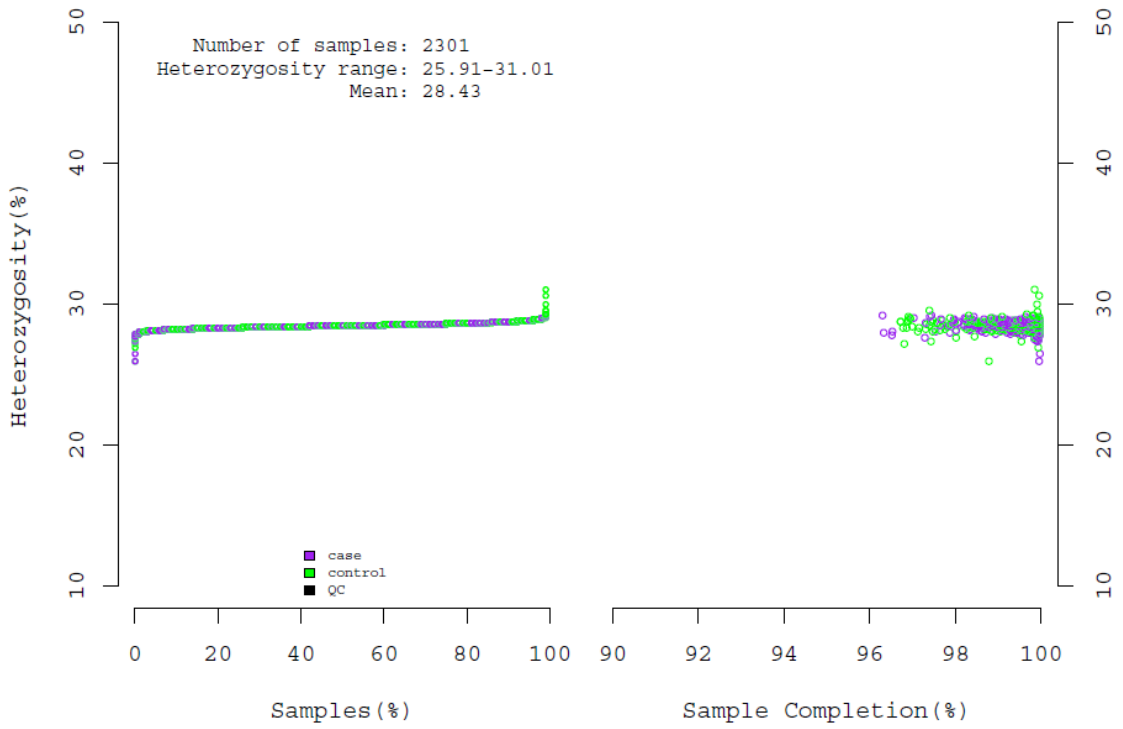

Figure 2b. Illumina 240K heterozygosity distribution (left) and its relation to completion (right).

2.3 Deviation from Hardy-Weinberg Proportions

Deviations from Hardy-Weinberg proportions (HWP) were assessed on the control samples for both assays. A quantile-quantile (Q-Q) plot of the p values for each group is shown in Figure 3a. Expected p values were calculated using the uniform distribution for all loci, and the observed p values were from an exact test for deviation from HWP[2]. SNPs were filtered by only including autosomal SNPs with MAF greater than 5%, and completion rates greater than 95%. The numbers of SNPs with p values less than 0.05 and 0.001 are shown in the plots. To better examine the lower tail of the p value distribution, Q-Q plots of the negative logarithm of the p values are shown in Figure 3b. In the low p-value region (<0.01) observed p-values are lower than the expected indicating that a mild deviation from HWP exists but does not pose problem to the overall genotype quality. Loci with p value lower than 10^{-7} were removed from the final analytic data set.

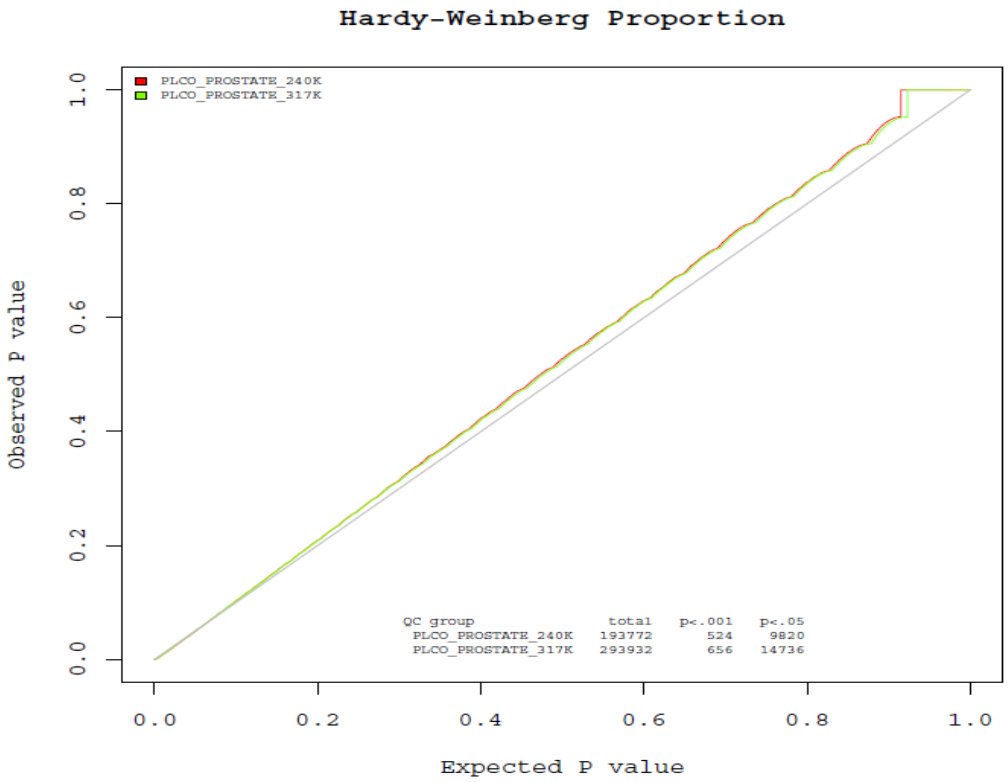

Figure 3a. Q-Q plot of tests for deviations from Hardy-Weinberg Proportions

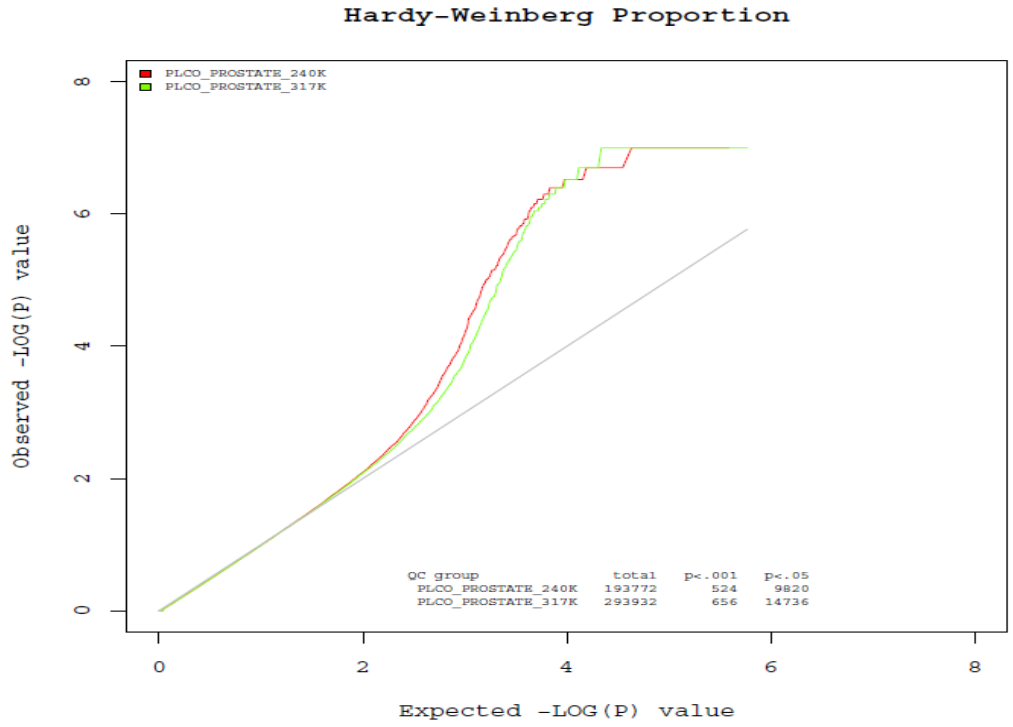

Figure 3b. Log-scale Q-Q plot of tests for deviations from Hardy-Weinberg Proportions

2.4 Assay Concordance

The assay concordance analysis was performed, and a total of 48 expected duplicates were identified in each QC group. The overall average concordance rate is 99.97%, and details are listed in the table below.

Table 3. Concordance for Expected Duplicates

Assay	Duplicate Pairs	Concordant Genotypes	Informative Comparisons	Concordance Rate (%)
Illumina 317K	48	8,091,567	8,094,060	99.97
Illumina 240K	48	6,215,407	6,217,209	99.97
Total	96	14,306,974	14,311,269	99.97

In addition, there are three pairs of unexpected duplicates with concordance rates greater than 99% in the QC group of Illumina 317K data. Three out of the 6 samples were also genotyped on the 240K chip. There are 9 assays in total classified with unclear identity and have been excluded from the final analytic data set.

2.5 Sex Verification

All study subjects self-reported as being male. Gender verification was performed by examining the heterozygosity of all loci genotyped on the X chromosome. The X chromosome mean sample heterozygosity is shown in Figure 4a and 4b. There are 3 individuals with data on the Illumina 317K assay with chromosome X heterozygosity around 35%, and are therefore deemed to be female and were not carried forward to be genotyped with the Illumina 240K assay.

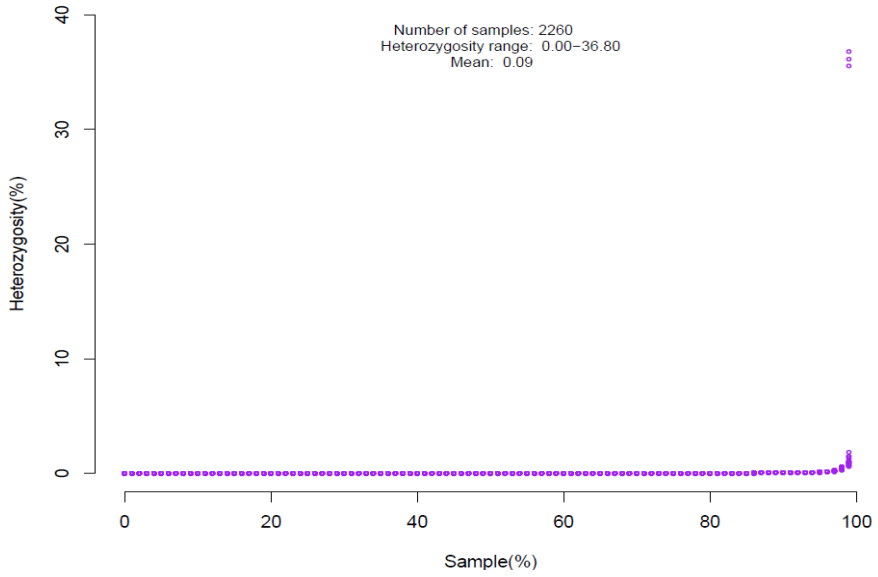

Figure 4a. Illumina 317K chromosome X heterozygosity distribution

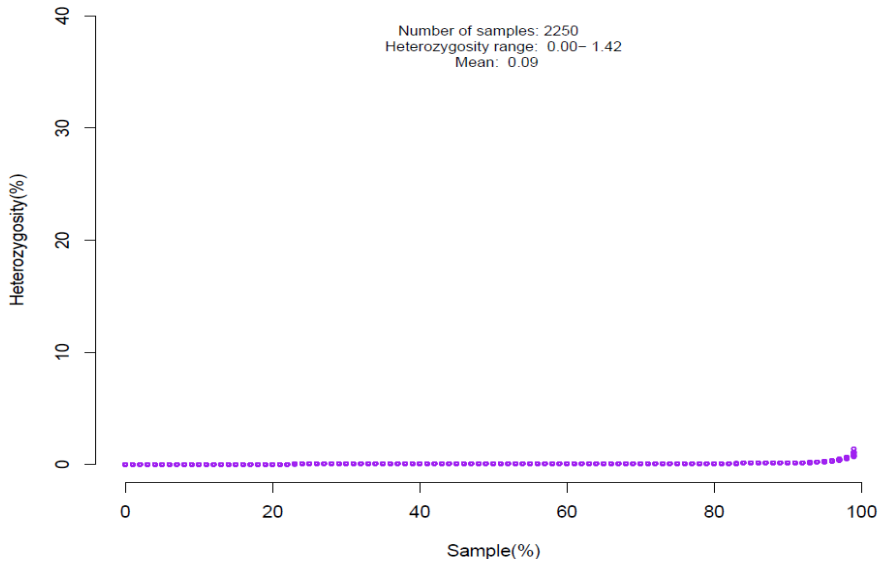

Figure 4b. Illumina 240K chromosome X heterozygosity distribution

2.6 Summary of Quality Control Exclusions

The distribution of sample missing rate and sample mean heterozygosity were examined for each QC group, and derived threshold for exclusion in each category is shown in table 4a. A total of 19 samples were excluded, and the details are listed in table 4b. In addition, a total of 7,203 loci were excluded, with details are listed in table 4c.

Table 4a. QC exclusion criteria

Criterion	Illumina 317K	Illumina 240K
Sample missing rate	>0.055	>0.04
Sample heterozygosity	<0.25 or >0.36	<0.25 or >0.35
Locus missing rate	>0.1	>0.1
HWP p value	$<10^{-7}$	$<10^{-7}$

Table 4b. Summary of excluded samples

Number of Samples	Illumina 317K	Illumina 240K
Sample missing rate	7	0
Sample heterozygosity	2	0
Unclear identity	6	3
Gender discordant	3	0
Total samples excluded*	16	3

* Counts reflect unique samples, as samples can be excluded based on multiple criteria.

Table 4c. Summary of excluded loci

Number of Samples	Illumina 317K	Illumina 240K
Locus missing rate	3,678	2,793
Extreme deviation from HWP	293	439
Total loci excluded	3,971	3,232

2.7 Relatedness Check

A check for relatedness was performed using the GLU *qc.ibds* module (http://code.google.com/p/glu-genetics/) to detect close relationships (1-2nd degree) using a method of moments estimator of allele sharing identical by descent. Five sibling pairs were found and were plausible: all are cases, and each pair is from the same study center.

2.8 Assessment of Ancestry and Population Structure

Ancestry was estimated for the 2,257 study subjects using a set of population informative SNPs (Kai Yu et al. PLoS ONE 2008) and data from the HapMap CEU, YRI, and ASA populations. These SNPs used are common to the commercially available Affymetrix 500K, Illumina 317K, and 550K chips. Admixture coefficients were estimated for each subject using the GLU *struct.admix* module (http://code.google.com/p/glu-genetics/), using the HapMap data as fixed reference populations. A total of 3 subjects (all controls) were estimated to be of less than 80% European ancestry, as shown in Figure 5.

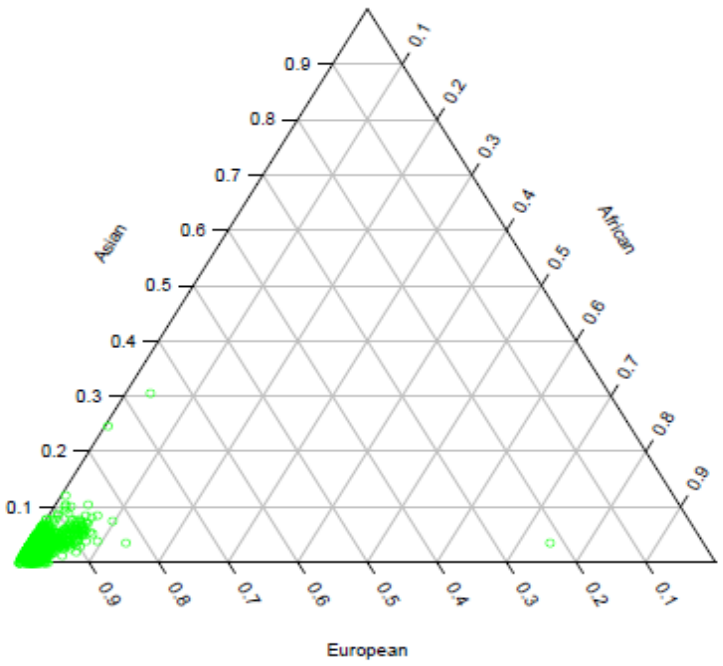

Figure 5. Population structure

To resolve more detailed differences in underlying population substructure, a principal component analysis was conducted on the final dataset with the same set of population informative SNPs described above using the GLU *struct.pca* module (http://code.google.com/p/glu-genetics/). Subjects with less than 80% European ancestry and one from each of the unexpected duplicates or relative pairs were excluded. Plots based on the first 6 principal components are shown in Figure 6.

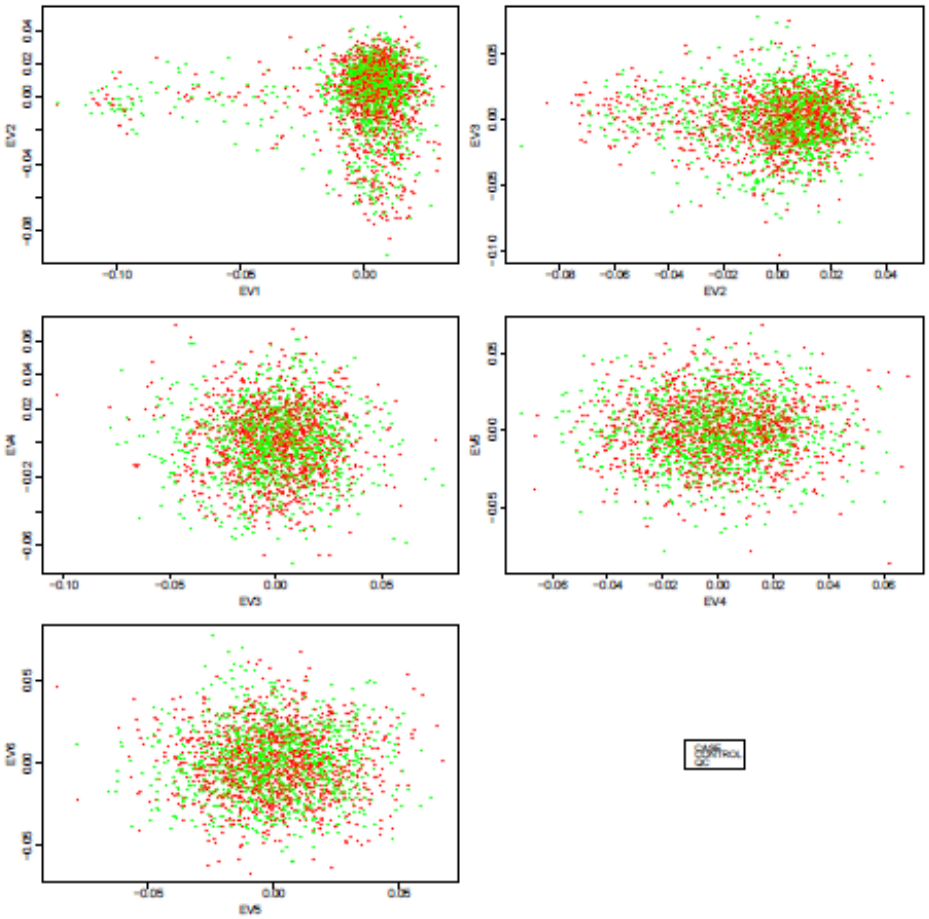

Figure 6. PCA plot (EV1-EV6).

Eigenvectors obtained from PCA can be used as covariates to adjust for the possible population stratification in the association model. Although we see evidence for slight differences in underlying substructure within our study population, these differences may not be potential confounders in genotype-phenotype association models. In order to evaluate this potential, we tested for correlation between eigenvectors with the case/control status by fitting a model including center, age, and the top 10 eigenvectors as independent variables to predict both case/control or aggressive/non-aggressive/control phenotype categories. As a result, the EV1 is significant with a Wald test p value less than 0.05, and has been included in the adjusted association models.

2.9 Analytical Exclusions

We excluded 3 admixed individuals as shown in section 2.8 and the only 2 additional subjects from a specific PLCO center. Final subject counts used in the association analysis are listed in table 6.

Table 6. Summary of subjects used in association

Phenotype	Subjects
Aggressive cases	659
Non-aggressive cases	492
Controls	1,101
Total	2,252

3 Differences from Build 1.0

3.1 Genotype Concordance

The genotype concordance rate between builds 1.0 and 2.0 was compared for subjects included in both builds. When counting all possible classes of discordances, including genotypes that were inconstantly missing between builds, the overall concordance rate is 99.85% for the Illumina 317K platform, and 99.35% for the Illumina 240K platform. The discordant genotypes are summarized in the table 7 and categorized by whether the discordant genotype was a homozygote, heterozygote or missing in each build. Counting only discordances between non-missing genotypes in both builds, a total of 26,574 genotypes were discordant out of 1,257,774,367 comparisons, a discordance rate of less than 2.12×10^{-5} with no homozygote genotypes converting to the opposite discordant homozygote (such discordances, if observed, would be evidence of extreme instability of genotype clustering). The vast majority of discordant genotypes are from genotypes present in Build 1.0 and excluded and set as missing from Build 2.0, a result of applying a genotype quality score cutoff based on the Illumina GC score of 0.25 that was not applied to build 1.0. The Illumina 240K assay exhibited a larger proportion of homozygote genotypes in Build 1.0 that were set to missing in Build 2.0. This is likely due to manual locus exclusions based on low clustering performance applied by Illumina Inc. to the Build 1.0 data.

Table 7. Summary of discordant genotypes between Build 1.0 and 2.0

		Illumina 317K assay *Build 2.0*			Illumina 240K assay *Build 2.0*		
		het.	hom.	missing	het.	hom.	missing
Build 1.0	**heterozygote**	*n/a*	297	229,795	*n/a*	5	115,358
	homozygote	71	0	791,620	26,201	0	3,394,595
	missing	32,121	34,964	*n/a*	8,275	6,769	*n/a*

3.2 Locus Comparison

The 554,291 loci passed genotype QC and are available for association analysis in build 2.0, compared to 546,593 for build1.0 (a gain of 7,698 loci). 10,090 loci are only present in build 2.0, and 2,391 loci are only present build 1.0. 544,202 SNPs (~99%) are present on both. The dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP) identifier (rs-number) for 5,264 loci were updated to conform to dbSNP build 130 and a table detailing the renaming applied will be distributed along with the data.

3.3 Sample Comparison

After QC excludes, a total of 4,603 samples remain in build 2.0 compared to 4,647 in build 1.0. Additionally, 44 samples (belonging to 25 subjects) including 37 ineligible samples and 7 failed the QC criteria were excluded in build 2.0. After applying the same analytic exclusions, the number of subjects that are used in association analysis is 2,252 for build 2.0, compared to 2,277 for build 1.0.

4 Association Analysis

We present results from two distinct analytic approaches. The first scheme is more frequently used in case control studies. The second scheme takes full advantage of the prospective nature of the PLCO cohort and the power from incidence density sampling.

Cumulative density sampling

For this scheme, which will be more familiar to non-epidemiologists, does not account for the dynamic nature of the cohort. Genotypes of individuals that have been selected as a case in the relevant phenotype case group are counted once as a case and never as a control. Individuals who have been selected several times as controls but had not developed prostate cancer during follow-up are counted only once in the control group.

Incidence density sampling

Selection of controls from cases identified in a cohort that accounts for the dynamic nature of the cohort including development of disease during follow-up and timing of entry to and exit from follow-up may have more power to detect an association than the single-selection method. The main feature of incidence-density sampling, as used for control selection here, is that controls are selected independently for each case among those who are at risk at the time of the diagnosis of the case; i.e., among those who would become a case in the study had they developed disease at the same time. Inclusion as a control for a given case set is independent of future diagnosis as a case, of selection as a control for other case sets, and of entry and exit times. Thus, individuals may be included as a case and as a control. Genotypes of individuals who have been selected multiples times are taken into account each time he is selected; the man's covariates that vary with time, such as age are defined differently each time, depending on the characteristics of the case set for which he was selected as a control[3].

The number of association model we fit increased from 4 in Build 1.0 to 32 in Build 2.0, including all combinations from the following four categories:

1. Sampling	
Cumulative density	Whole genome association analysis of main effects for 554,291 SNPs on 1,151 cases diagnosed with tumors and 1,101 controls that were not diagnosed with prostate cancer at the start of the CGEMS project.
Incidence density	Whole genome association analysis of main effects for 554,291 SNPs on 1,151 cases diagnosed with tumors and 1,156 controls selected using an incidence density sampling strategy.

2. Dependent variable in model	
Dichotomous	A dichotomous logistic model was constructed to contrast the risk of all prostate cancer cases (both non-aggressive and aggressive) against that of all controls (m=2).
Polytomous	A polytomous logistic model was constructed to separately contrast the risk of non-aggressive and aggressive prostate cancer cases against that of all controls (m=3).

3. Covariate adjustment	
Unadjusted	A 3-by-m contingency table of genotypes by phenotypes was constructed.
Adjusted	The m phenotypes were regressed on indicator variables for genotype effects, age group at randomization (4 groups), region of recruitment (9 non-reference regions), and a single eigenvector to account for population stratification.

4. Genotype effects	
Genotypic	The p-value was obtained from a score test of each estimated genotype effect with up to 2(m-1) degrees of freedom. (m is the number of phenotype categories)
Trend	The p-value was obtained from a score test for the estimated trend of the genotype effects with up to m-1 degrees of freedom.
Dominant	The p-value was obtained from a score test for the minor homozygote + heterozygote versus major homozygote effect with up to m-1 degrees of freedom.
Recessive	The p-value was obtained from a score test for the minor homozygote versus heterozygote + major homozygote effect with up to m-1 degrees of freedom.

The GLU *assoc.logit1* module (http://code.google.com/p/glu-genetics/) was used to fit all models and to perform score tests of all genetic terms for association with phenotype.

References

1 Yeager, M. *et al*. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39**, 645–649 (2007).

2 Wigginton,J.E., Cutler,D.J. & Abecasis,G.R. A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 887-893 (2005).

3 Wacholder,S., Silverman,D.T., McLaughlin,J.K. & Mandel,J.S. Selection of controls in case-control studies. III. Design options. *Am. J. Epidemiol.* **135**, 1042-1050 (1992).